Classification and retrieval of CERES cloudy and clear scenes from TOA radiances using Random Forests method

Bijoy Vengasseril Thampi[1], Constantine Lukashin[2]
Takmeng Wong[2]

[1]Science System Applications Inc., Hampton, VA
[2]NASA Langley Research Center, VA

Joint CERES/GERB/ScaRaB Science meeting, Toulouse, 7-10 October 2014

Background

- CERES TOA radiances are converted to TOA fluxes using empirical angular distribution models (ADMs).
- However, in the absence of imager coverage over the CERES footprints or unavailability of imager data (e.g., due to malfunction of the instrument), accurate scene identification and subsequent estimation of TOA fluxes are difficult.
- It is observed that 5.6% of all CERES Terra/Aqua footprints contains missing imager information or insufficient imager data for a reliable scene ID and it can reach up to 50% of data for a specific scene types.
- The Big question: how cloud/clear scene determination is to be carried out in case of imager data become unavailable..

Objective

The motivation for this study is to develop a methodology for the improved estimation of scene type (clear/cloudy) using CERES TOA radiances and other ancillary measurements without using any imager data.

Approach: Use ensemble classifier like “Random Forests“ to derive the cloud condition

Random Forests

- ***Random Forests (RF)***, is an ensemble classifier similar to ANN, SVM, Naive Bayes..
- RF generate ***multiple learning classifiers and aggregate their results*** to obtain better predictive performance than could be obtained from single classifier.
- Use **decision tree classifiers** as the base classifier.
- Main advantages of RF method are

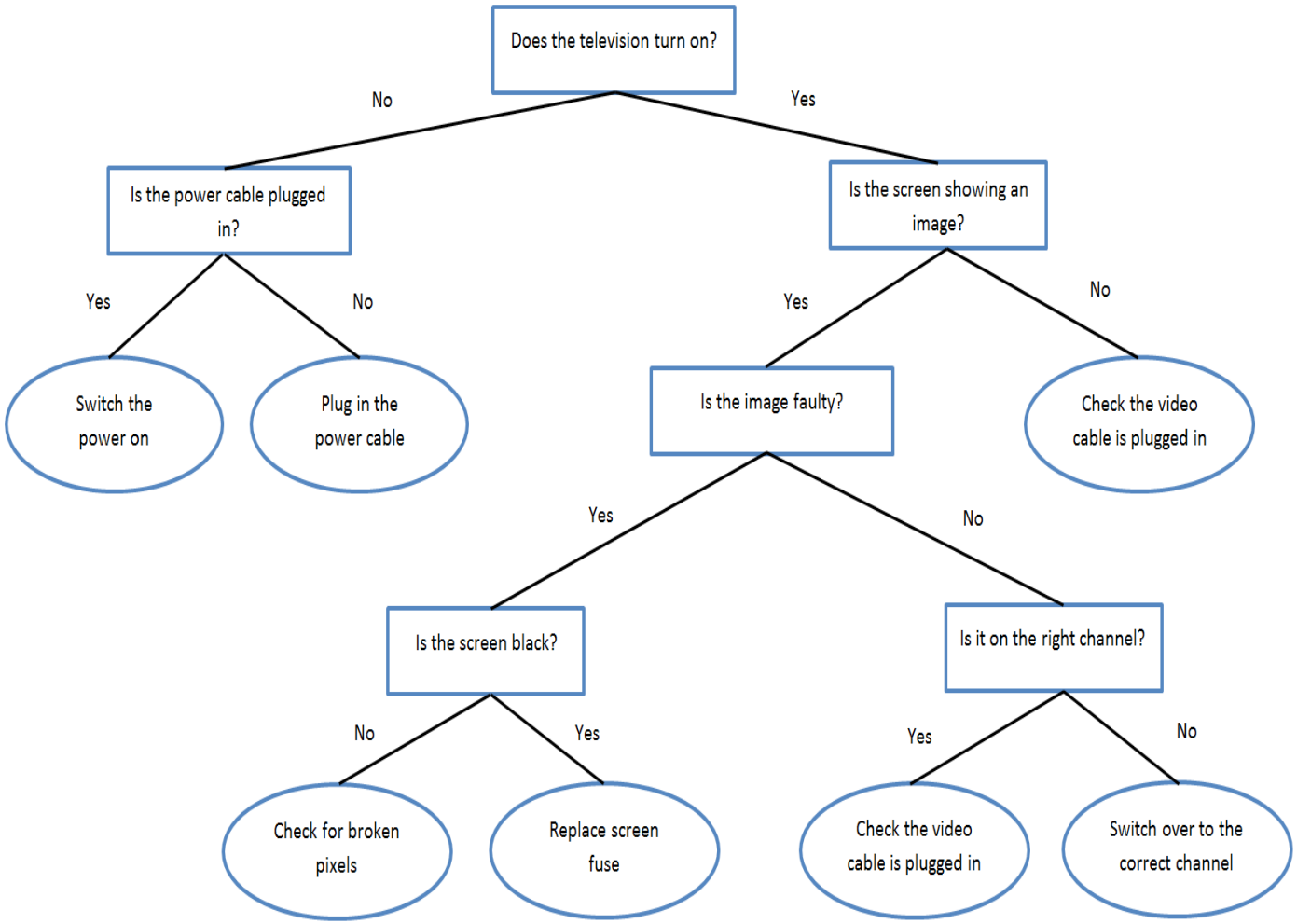
i) they have faster run times

ii) they can deal with unbalanced and missing data

iii) ability to handle data without preprocessing or rescaling.

Decision Tree

Decision tree predictors are the basic unit of random forest. Simple decision trees are appealing because of their clear depiction of how a few inputs can determine the output.



Random Forest Algorithm

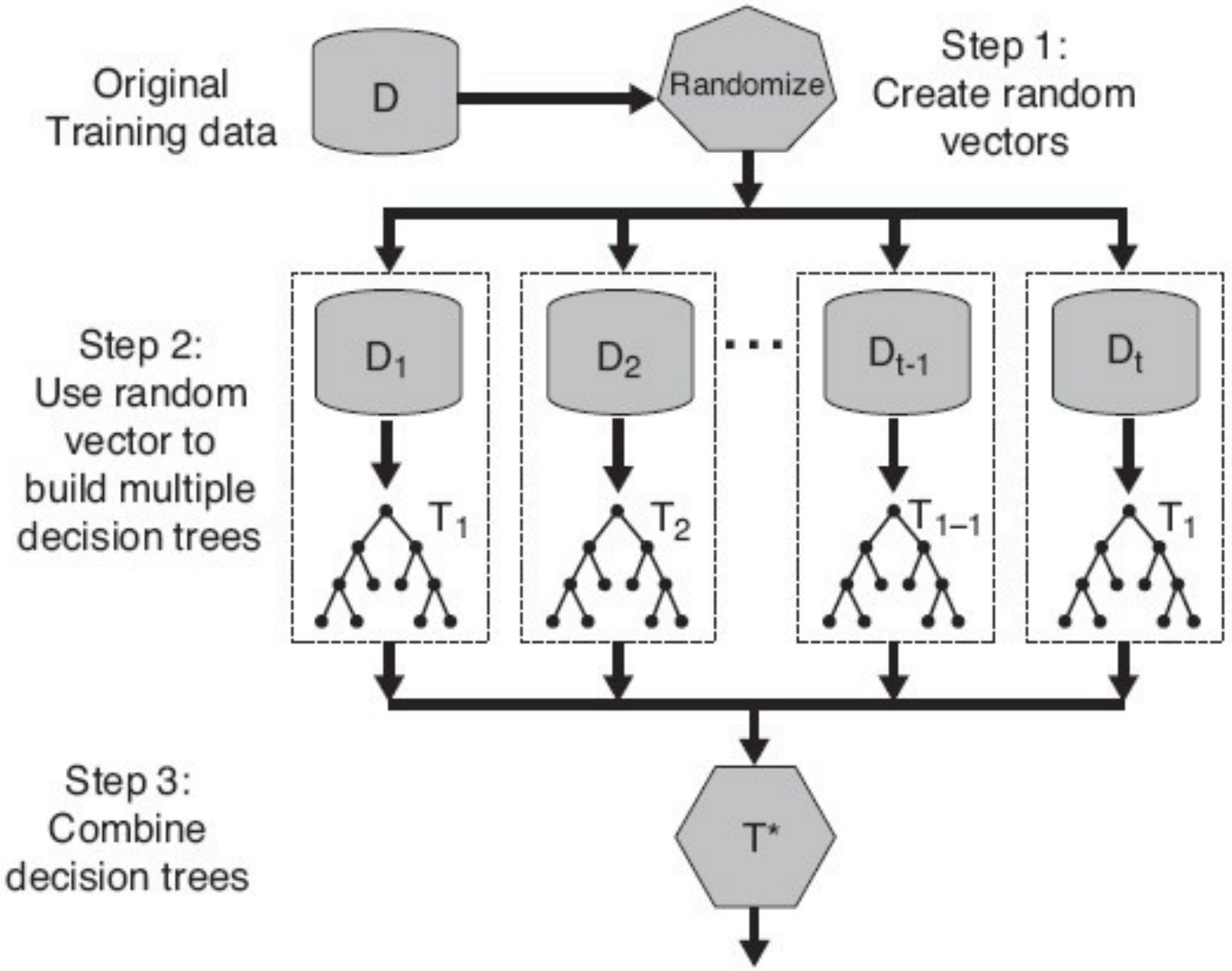
- Introduce two sources of randomness: “**Bagging**” and “**Random selection of input vectors**”.

Bagging- repeated random sub-sampling of the training data.

Bootstrap sample - will on average contain **63.2%** of the data while the rest are replicates.

- Using bootstrap sample, a decision tree is grown to its greatest depth using the training data.
- For each tree, using the leftover (**36.8%**) data, misclassification rate is calculated (**out of bag (OOB)** error)
- Aggregate error from all trees to determine **overall OOB error rate for the classification**

Random Forests –Flow diagram



Input Variables

Input variables are selected for the scene classification are:

CERES	Ancillary data
Solar zenith & viewing zenith angles Relative azimuth angle CERES TOA LW & SW broadband radiances IGBP Surface types	LW surface emissivity Broadband surface albedo Surface skin temperature Precipitable water

IGBP Surface Types	
Water bodies Bright Desert Dark Desert Grasslands Croplands and cities	Evergreen Forests Deciduous Forests Savannas and Shrublands Permanent and Fresh snow Sea Ice

Training data for RF Analysis

Training dataset is constructed by stratifying the data in the variable of interest and using the corresponding average class values in the RF analysis. Due to lack of data for some surface types, in the present analysis, **Aqua SSF monthly data from 2002 to 2012 period is used for the constrcution of training data set.**

Test data is constrcucted using subsampling of the monthly SSF footpring data.

Variables

Variable	Bin width	No. of Bins
SZA	1°	90
VZA	1°	70
RZA	1°	180
SWR (D)	20-40 $W/m^2/sr$	4-7
LWR (DN)	10-20 $W/m^2/sr$	4-7

RF analysis - Confusion matrix

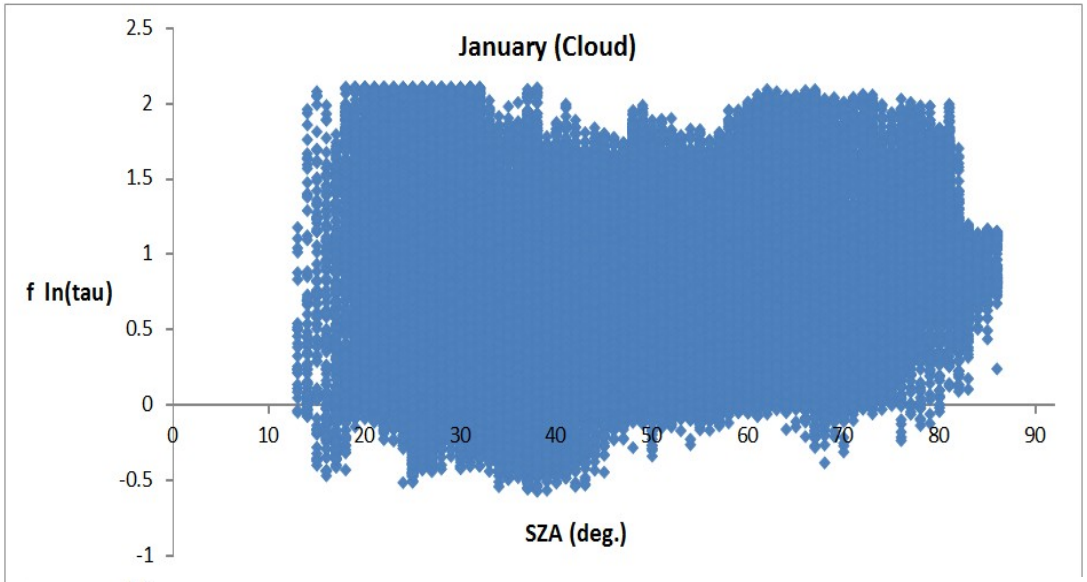
confusion matrix allows visualization of the performance of RF algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class

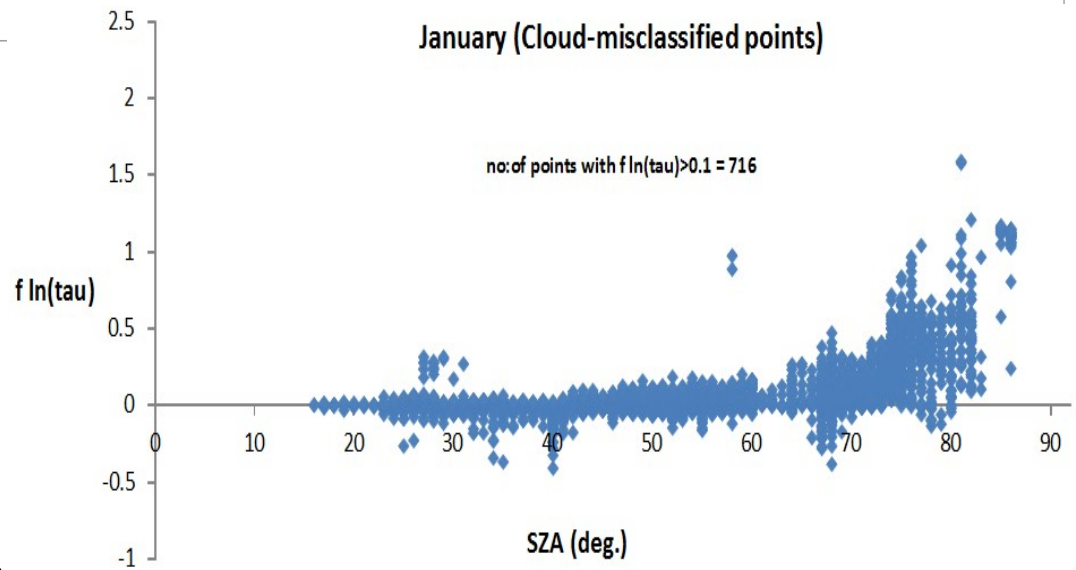
clear

Class	1	2	3	4	5	6	7	8	9	10	11	12
1	**62374**	36	0	0	0	10814	1	0	0	0	0	0
2	3	**22874**	5	0	0	0	5921	0	0	0	0	0
3	0	0	**29**	0	0	0	0	7	0	0	0	0
4	0	0	0	**7**	0	0	0	0	1	0	0	0
5	0	0	0	0	**0**	0	0	0	0	0	0	0
6	11569	30	0	0	0	**16139**	25	0	0	0	0	0
7	8	2950	0	0	0	103	**24301**	7	0	0	0	0
8	0	0	49	1	0	0	263	**11204**	20	0	0	0
9	0	0	0	28	0	0	0	59	**6558**	5	0	0
10	0	0	0	0	0	0	0	0	4	**5125**	0	0
11	0	0	0	0	0	0	0	0	0	22	**5762**	3
12	0	0	0	0	0	0	0	0	0	0	0	**14656**

Classified cloudy

RF analysis (Day time)





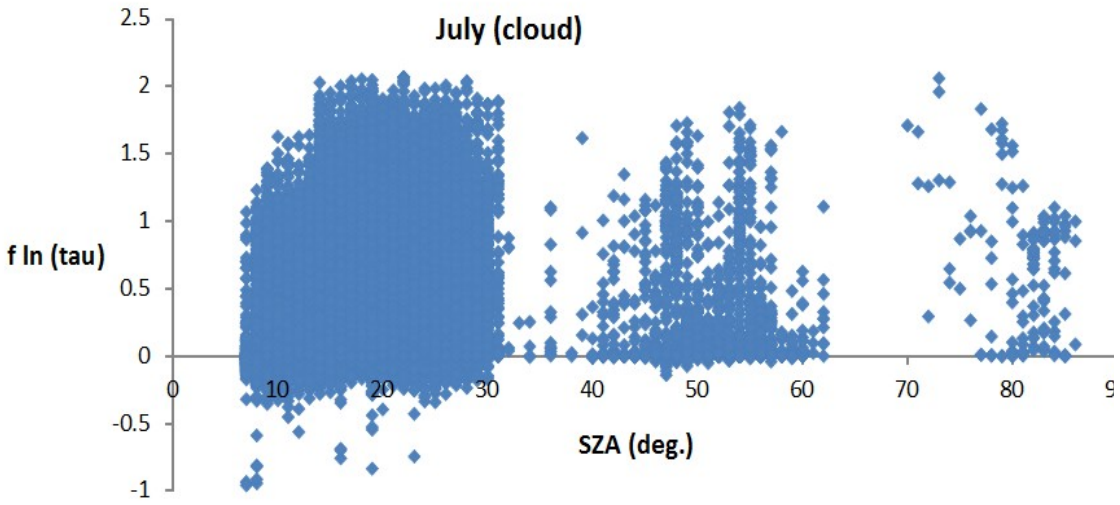
Surface type : Water bodies
Month : January

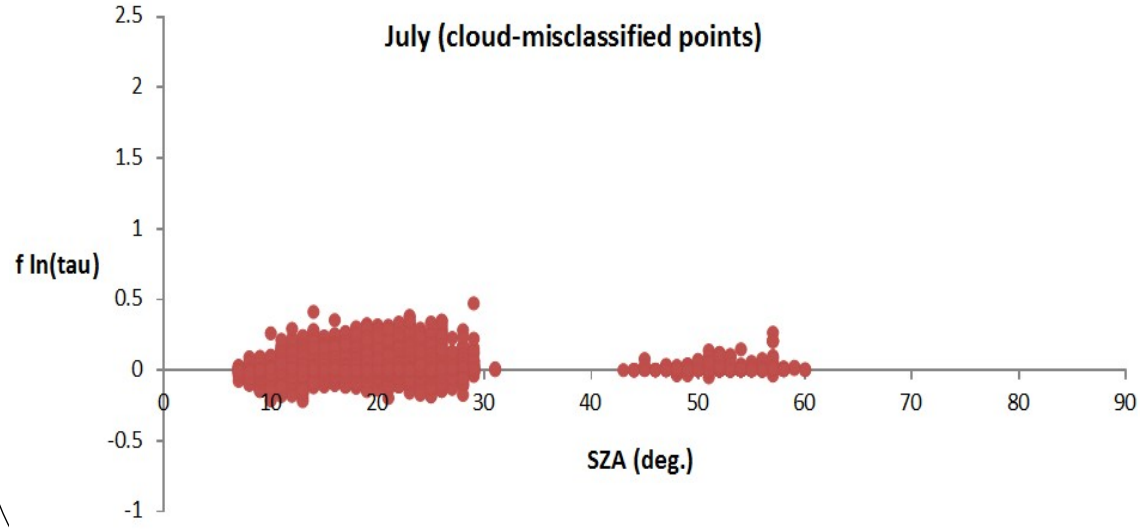
Number of test data points
- 100000 (only cloudy sky)

Total no. of Misclassified points
- 7987 (~8%)

% of Misclassified points with **f ln(tau)** >0.1 is = **0.72 %**

RF analysis (Day time)





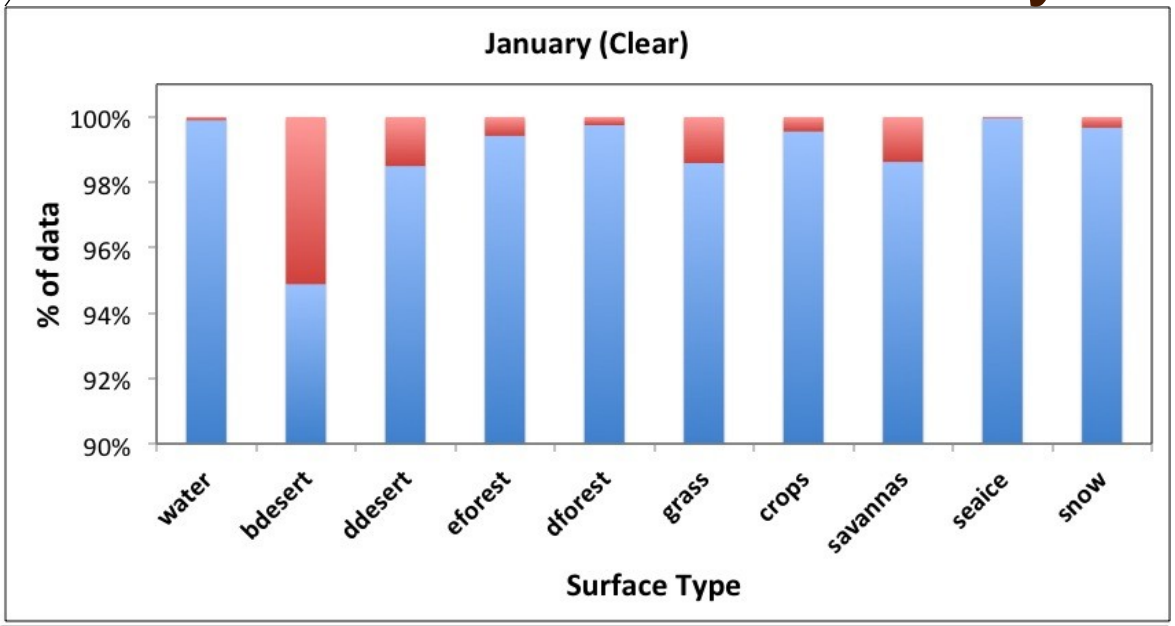
Surface type : Bright Desert
Month : July

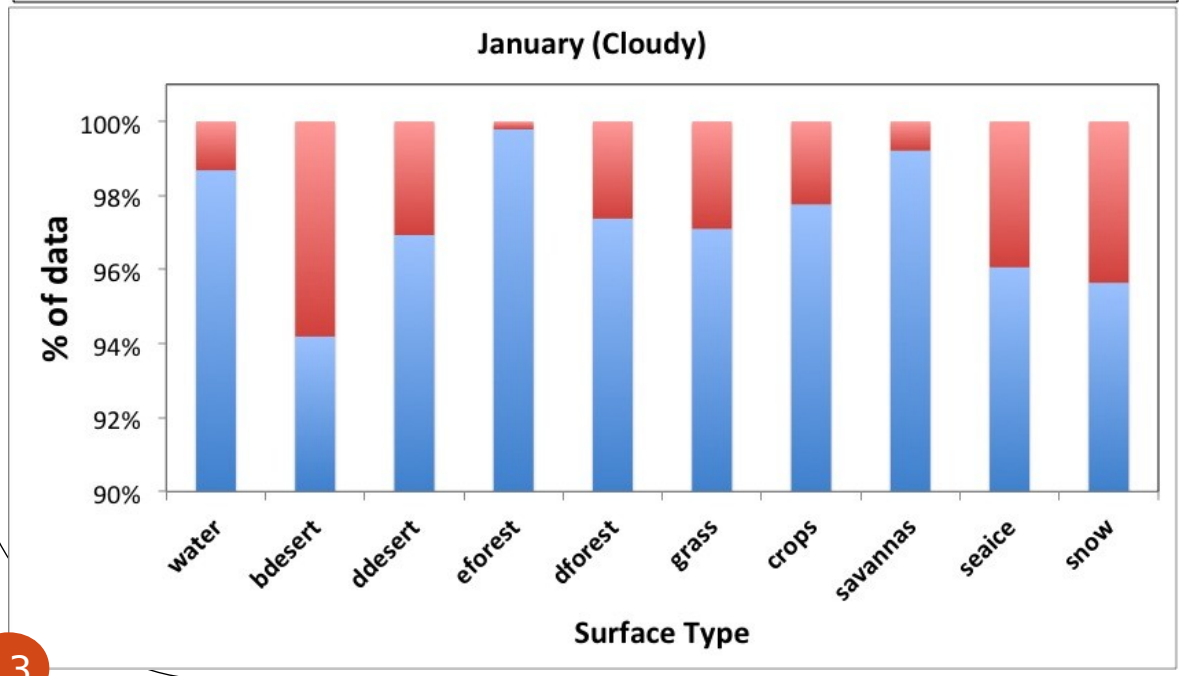
Number of test data points
- 100000 (only cloudy sky)

Total no. of Misclassified points
- 21649 (~21.7%)

% of Misclassified points with **f ln(tau)** >0.1 is = **0.92 %**

Error analysis





Month: January (Day time)

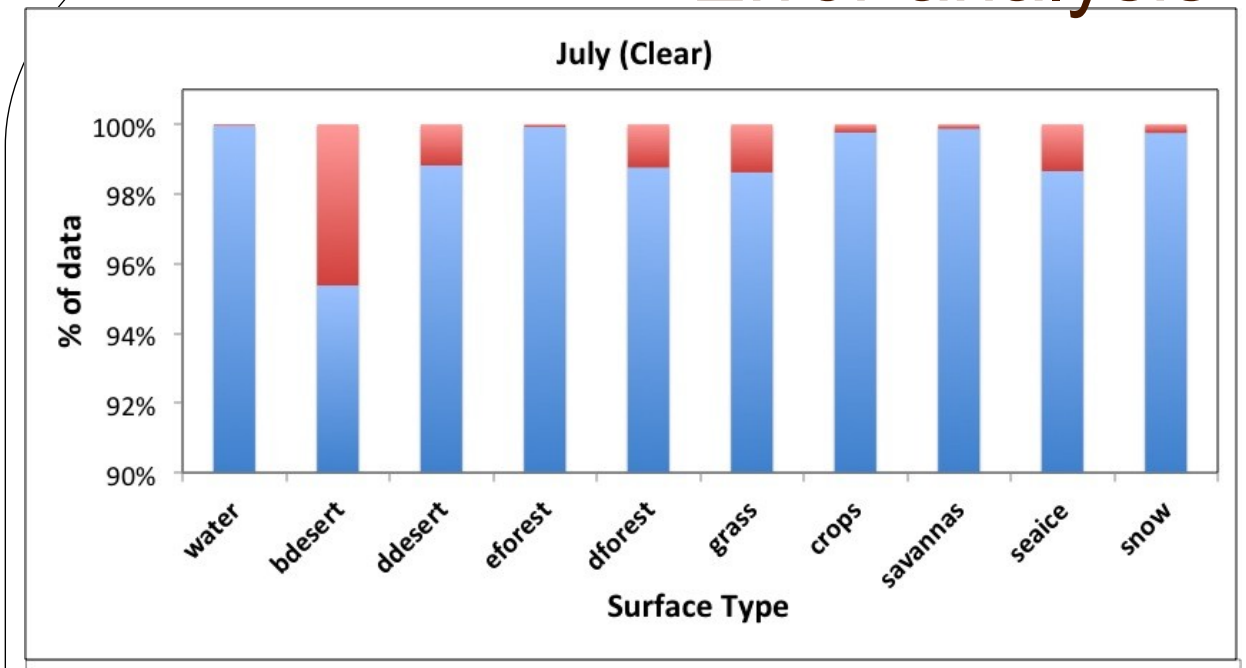
$diff = Rad_{training\ dataset} - Rad_{test\ dataset}$

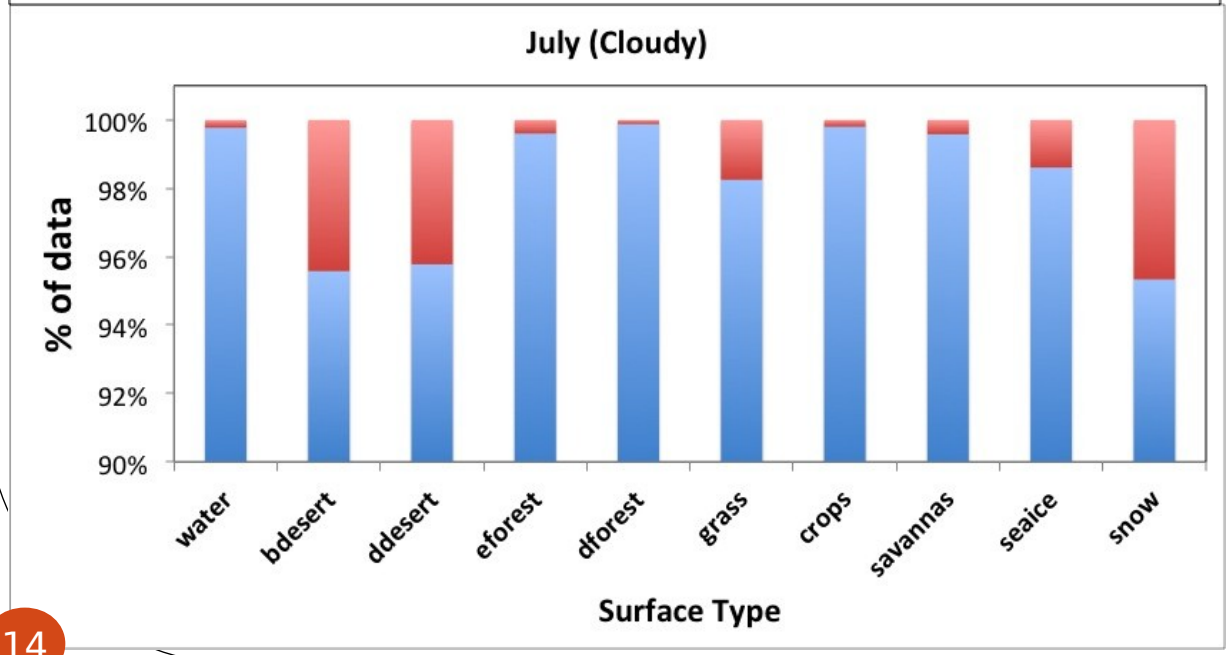
red = % of misclassified points with diff.>10 W/m^2/sr for each group

Relatively large misclassification rate was observed for surface types: Bright Deserts, Dark deserts, Grass lands and Snow

Lower misclassification rate was observed for surface types: Water bodies, Evergreen forest and Deciduous forests

Error analysis





Month : July (Day time)

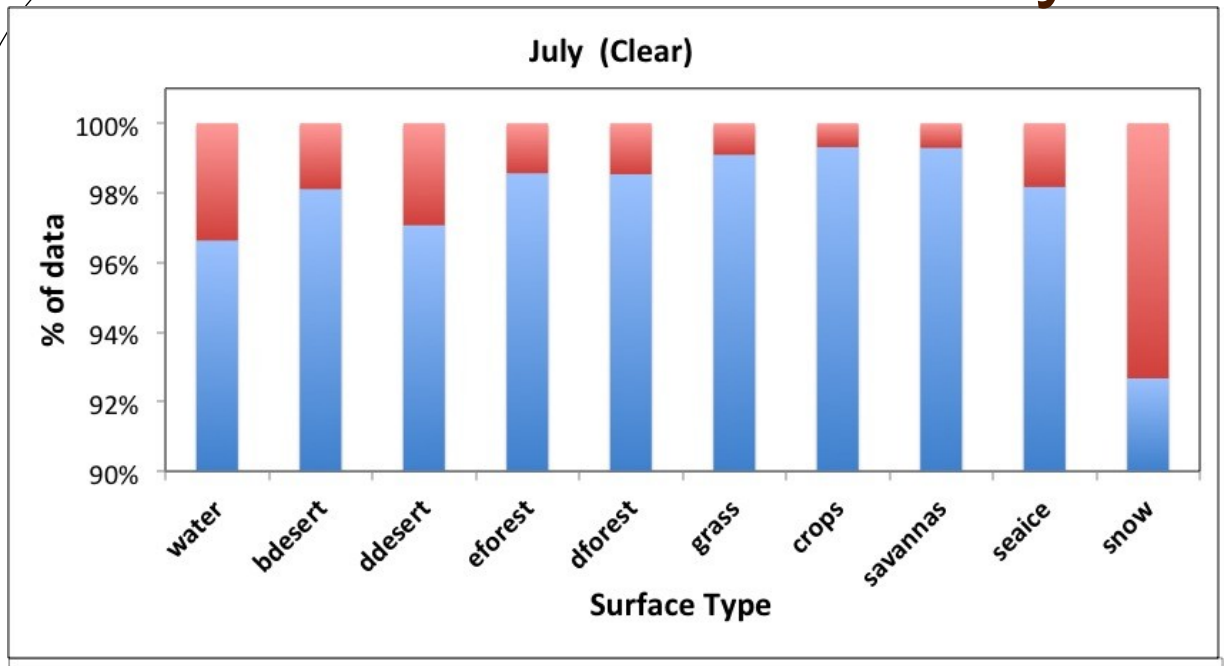
$diff = Rad_{training\ dataset} - Rad_{test\ dataset}$

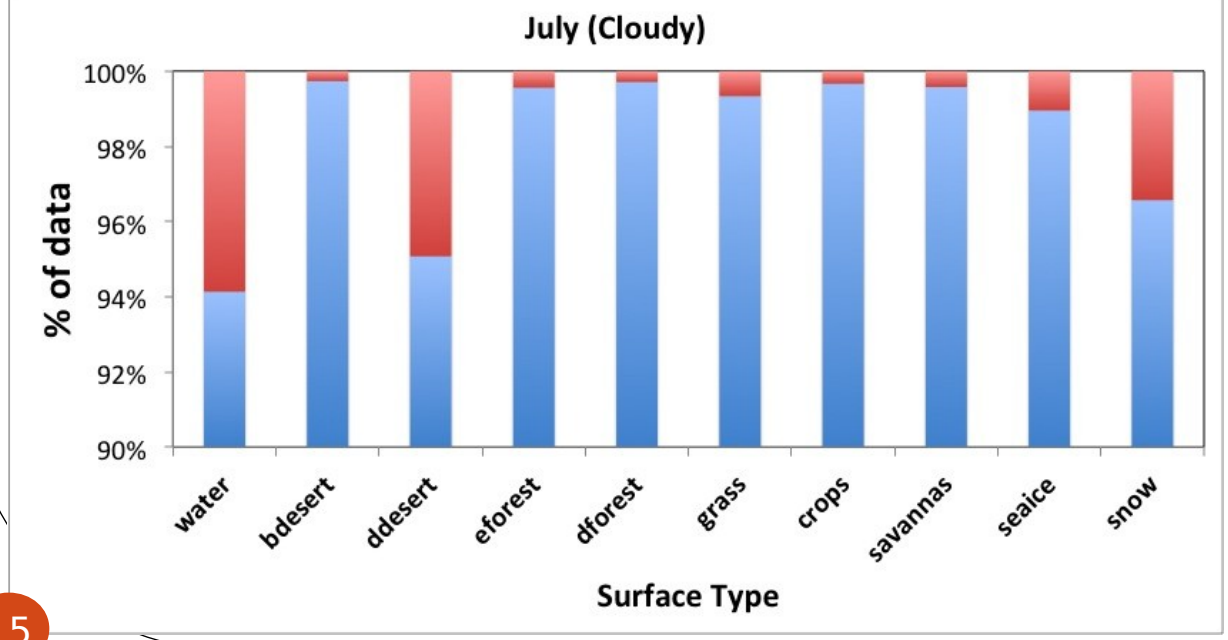
red = % of misclassified points with diff.>10 W/m^2/sr for each group

Relatively large misclassification rate was observed for surface types: Bright Deserts, Dark deserts, Sea ice and Snow

Lower misclassification rate was observed for surface types: Water bodies, Evergreen forest, Deciduous forests, Crops and Savannas

Error analysis





Month : July (Night time)

$diff = Rad_{training\ dataset} - Rad_{test\ dataset}$

red = % of misclassified points with diff.>10 W/m^2/sr for each group

Relatively large misclassification rate was observed for surface types: Water bodies, Dark deserts and Snow.

Lower misclassification rate was observed for surface types: Deciduous and Ever green forests, Crops, Seaice, Grasslands and Savannas

Conclusions

- RF scene classification using CERES TOA radiances show very good results for both day and night time
- RF misclassification rate for (Clear and cloudy, Day time) shows relatively lower values (misclassification rate < 2 %) for Water bodies, Crops, Evergreen forest, etc.,
- RF misclassification rate for (Clear and cloudy, Day time) shows large values (misclassification rate ~ 3-8 %) for Bright and dark deserts, Snow and Grasslands.

- **Future work:** Incorporation of output from RF analysis in to the ANN based estimation of TOA flux.

Thank you ….

Confusion matrix –test data

a RF confusion matrix allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class

eg., Month – July; surface type – water bodies.

	62374	36	0	0	0	10814	1	0	0	0	0	0	
classified clear	3	22874	5	0	0	0	5921	0	0	0	0	0	Misclassified clear
	0	0	29	0	0	0	0	7	0	0	0	0	
	0	0	0	7	0	0	0	0	1	0	0	0	
	0	0	0	0	0	0	0	0	0	0	0	0	
	11569	30	0	0	0	16139	25	0	0	0	0	0	
Misclassified cloudy	8	2950	0	0	0	103	24301	7	0	0	0	0	Classified cloudy
	0	0	49	1	0	0	263	11204	20	0	0	0	
	0	0	0	28	0	0	0	59	6558	5	0	0	
	0	0	0	0	0	0	0	0	4	5125	0	0	
	0	0	0	0	0	0	0	0	0	22	5762	3	

Machine learning

Machine learning deals with the construction and study of systems that can learn from data. It focuses on prediction, based on known properties of the system learned from the training data.

Reduced variance: results are less dependant on peculiarities of a single training set.

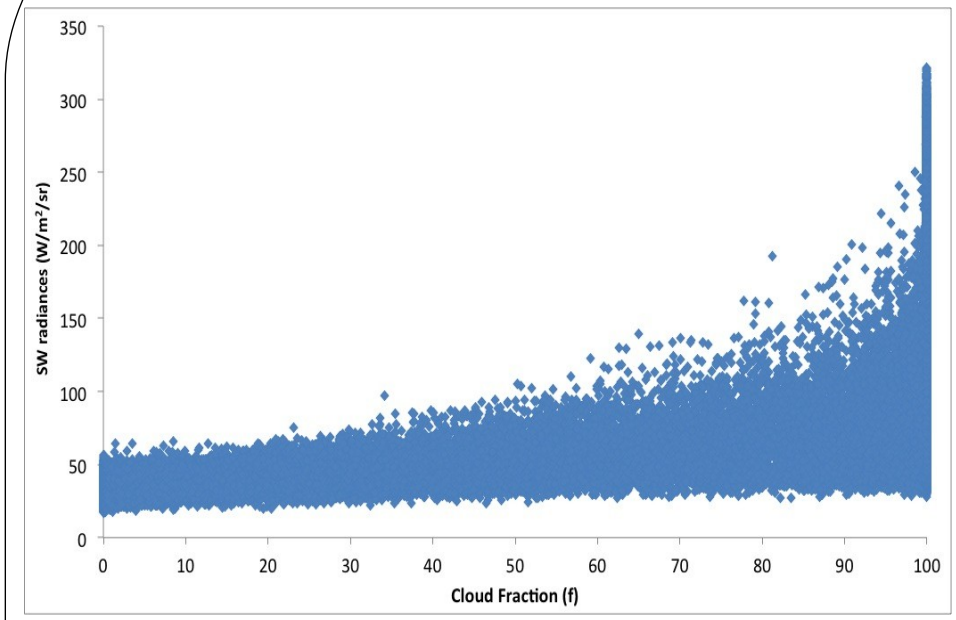
Reduced bias : combination of multiple classifiers may produce more reliable

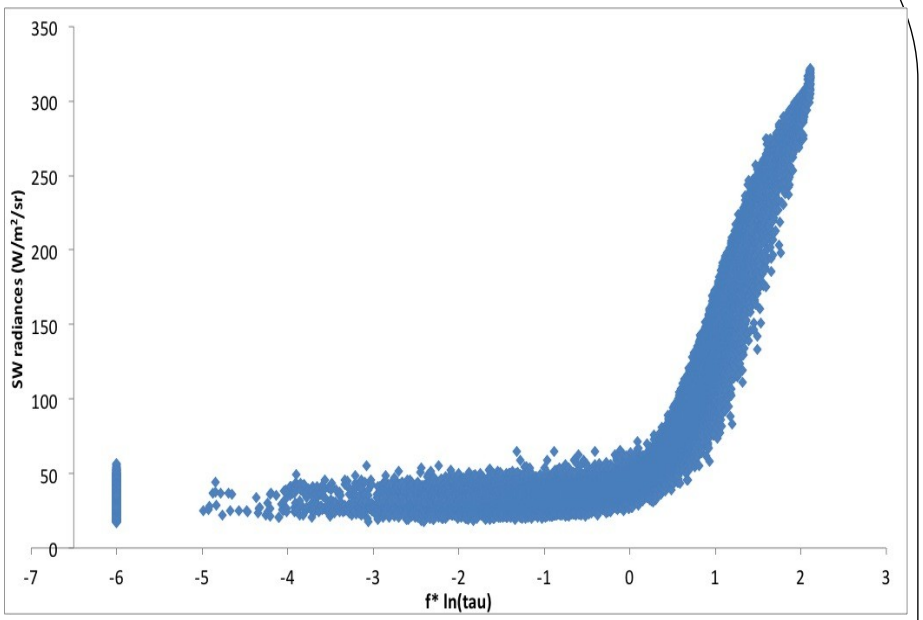
classification than single classifier.

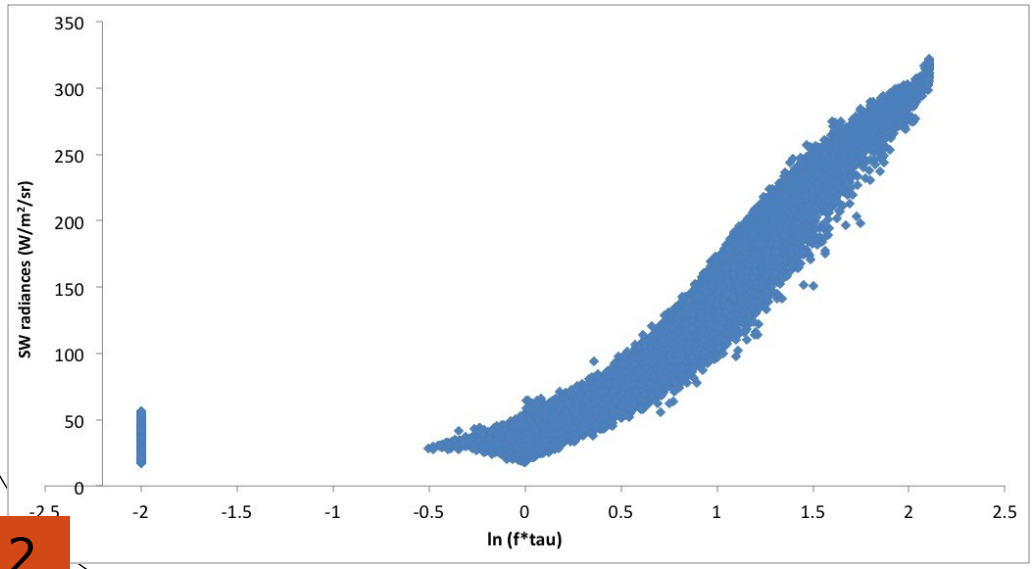
Eg., SVM, Bayes optimal classifier, Boosting, Bagging, Random forest

Random forests first proposed by Tin Kam Ho of Bell Labs

Building Training data set



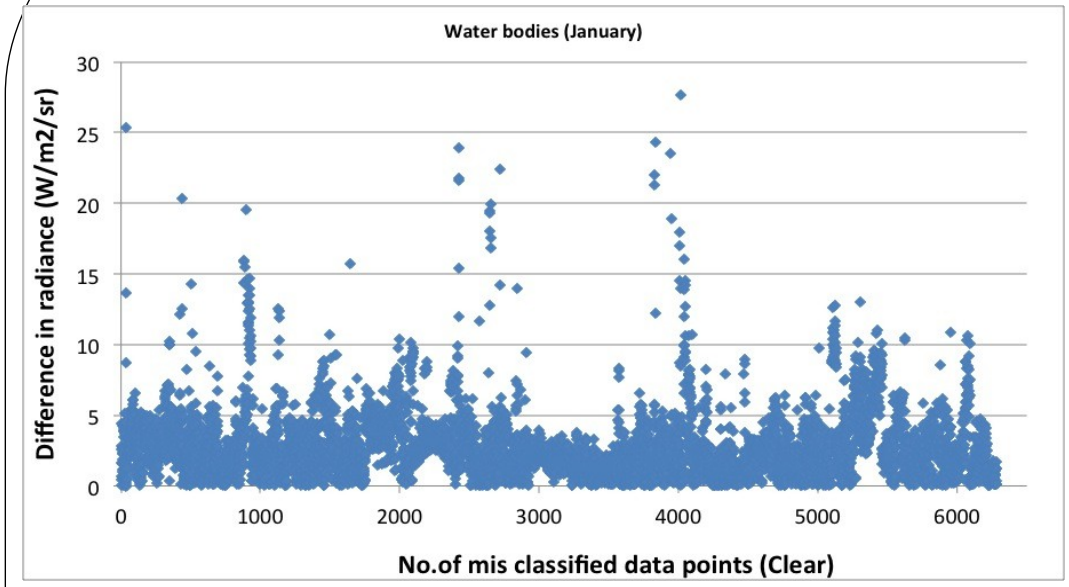


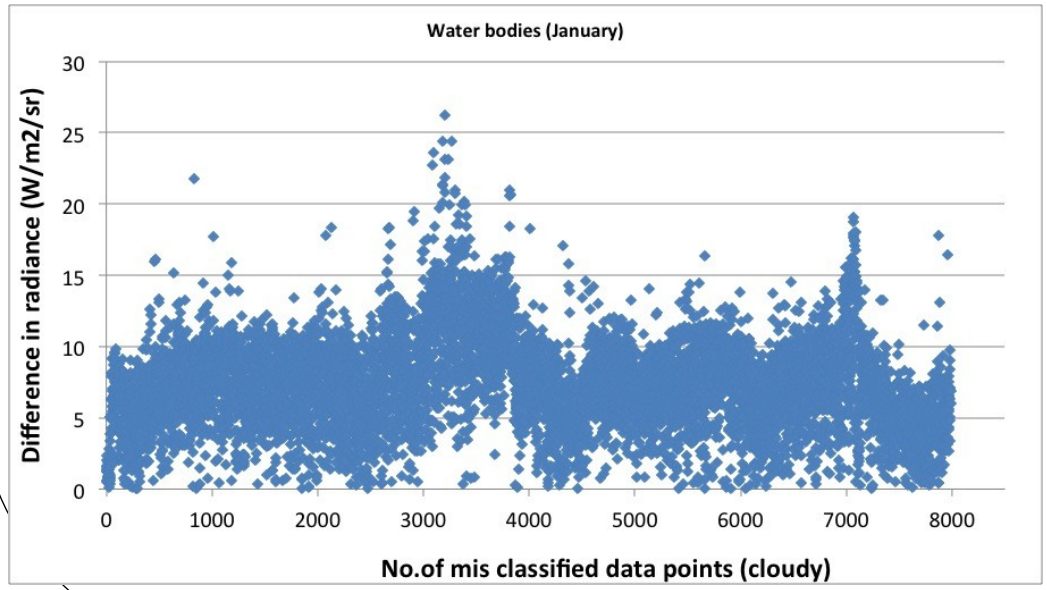


CERES SSF radiance values for the angular bins SZA=19-20, VZA=5-10.

Magnitude of CERES radiance in an angular bin is sensitive to the cloud fraction and cloud optical depth.

RF analysis- best case (Day time)





Surface type : Water bodies
Month : January

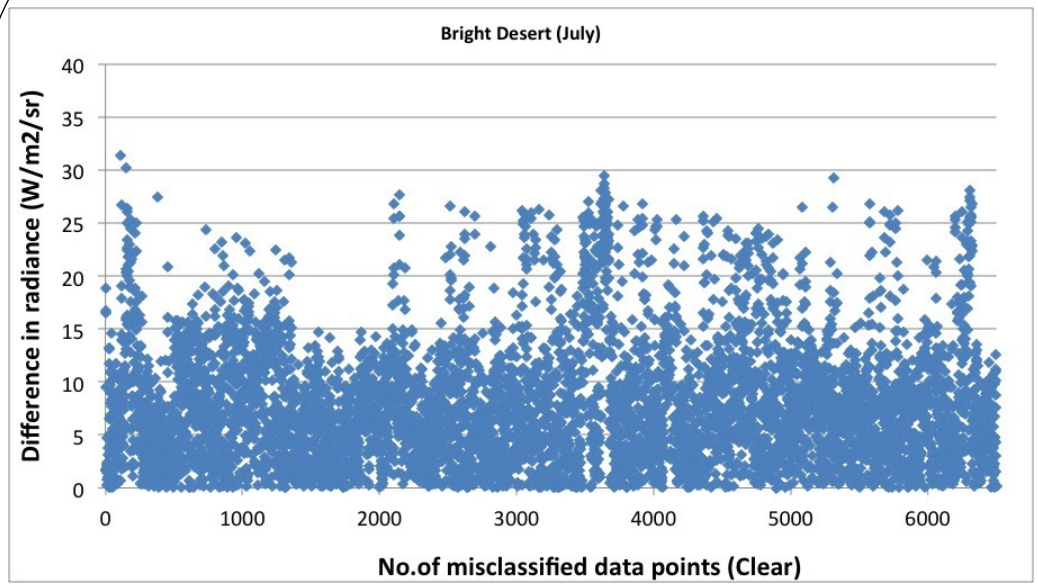
Number of test data points
Group 1 - 100000 (only clear sky)
Group 2 - 100000 (only cloudy sky)

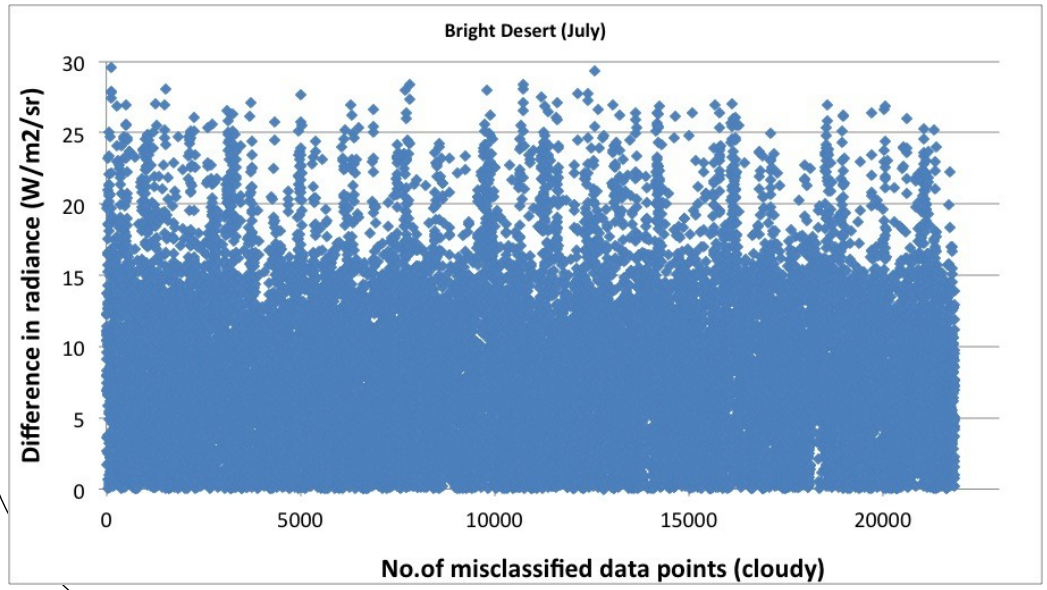
Total no. of Misclassified points
Group 1 - 6283 (6.2 %)
Group 2 - 7987 (~8%)

diff = training dataset values – test dataset value
% error = No. of Misclassified points with diff.>10 W/m2/sr for each group

Group 1 - 0.1 %
Group 2 - 1.33 %

RF analysis- worst case (Day time)





Surface type: Bright Desert
Month : July

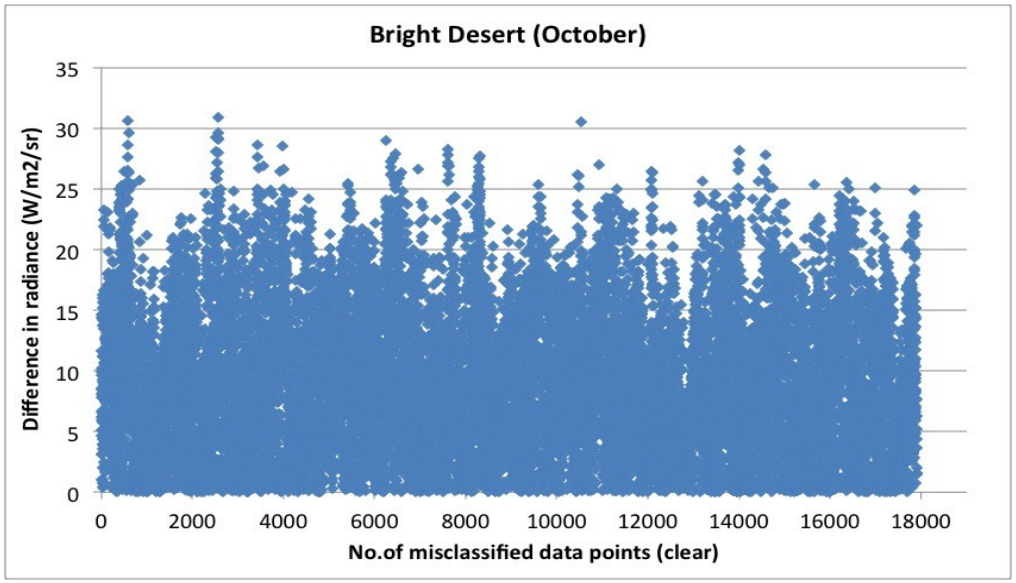
Number of test data points
Group 1 - 100000 (only clear sky data)
Group 2 - 100000 (only cloudy sky data)

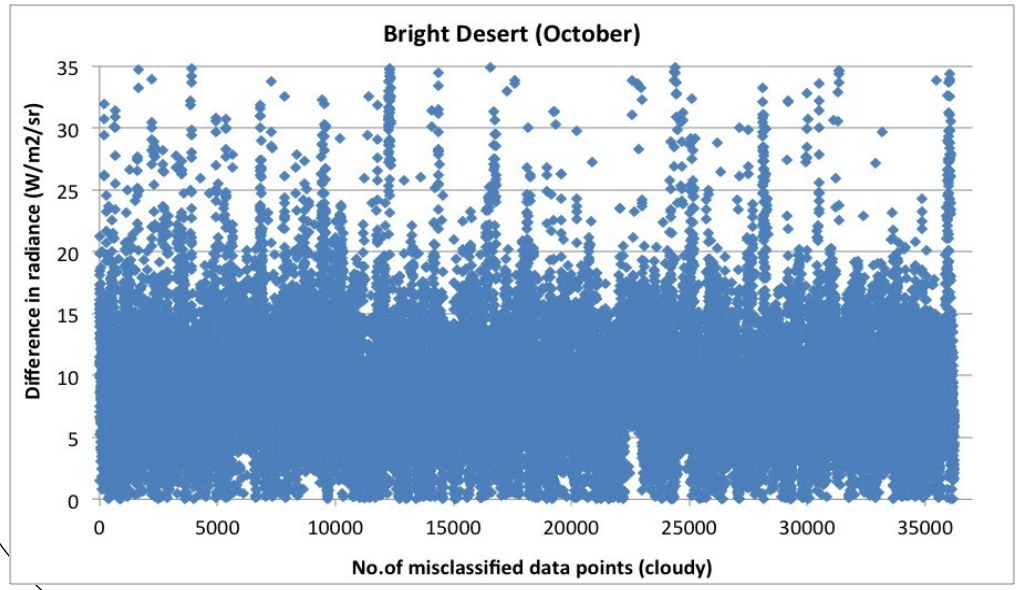
Total no. of Misclassified points
Group 1 - 24750 (~24.7%)
Group 2 - 21856 (~21.9%)

% error = No. of Misclassified points with diff.>10 W/m2/sr for each group

Group 1 - 6.65 %
Group 2 - 7.38 %

RF analysis- worst case (Day time)





Surface type : Bright Desert
Month : October

Number of test data points
Group 1 - 100000 (only clear sky)
Group 2 - 100000 (only cloudy sky)

Total no. of Misclassified points
Group 1 - 17903 (~17.9%)
Group 2 - 36225 (~36.3%)

diff = training dataset values – test dataset value
No. of Misclassified points with diff.>10 W/m2/sr. Number of data points misclassified as,

Clear - 7171
Cloudy - 11954